ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Voice Interchange

Vayusutha M [1], Dr. Ganesh V Bhat [2]

Asst. Professor, Department of EC, Canara Engineering College, Bantwal, Mangalore, India [1]

Professor, Department of EC, Canara Engineering College, Bantwal, Mangalore, India [2]

ABSTRACT: An idea of voice Interchange using Source Filter Model of speech Synthesis using Linear predictive coefficients (LPC) as Vocal Tract Parameter is proposed. Experimental result shows that Excitation parameter is related to content of the speaker where as vocal tract parameter is related to speaker style. Assessment of interchanged voice is made on both objective and subjective basis which convinces successfulness of interchange and obtained good transformation result in terms of formant frequencies. Utterances for the proposed work are chosen from CMU_ARCTIC speech synthesis database.

KEYWORDS: Voice Interchange, Voice Conversion, DTW, Source Filter Model

I. INTRODUCTION

Voice interchange is the process of interchanging the style and content parameter of speech between source speaker and target speaker respectively. Proposed work on Voice interchange provides an idea on how voice of source speaker can be modified to hear as a target speaker which in turn can be called as voice conversion [5]. One way to look at voice conversion is modifying one or more parameters of a source speaker in such a way that source speaker sounds like some other (target) speaker. Even though there is lot of research going on in voice conversion techniques, proposed work gives an idea of voice conversion for beginner.

One of the main applications of voice conversion is to create personalized voices in text to speech (TTS) synthesis. Other entertainment related applications include voice dubbing where any trained speaker can speak in another famous speaker voice or can be extended in such a way that any professional singer can sing in some other famous singer voice.

II. RELATED WORK

Literature survey reveals that work on voice conversion started from early 1970's to till date in which one of the initial works in modifying voice characteristics was done by Atal and Hanauer (1971) [1], in this work authors worked on representing voice as transfer function of vocal tract and characteristics of excitation and also studied the feasibility of modifying voice characteristics using LPC vocoder. Seneff (1982) [2], demonstrated a method to modify the excitation and vocal tract parameters. The trend of voice conversion began in the year 1985 where childers et al, (1985, 1987, 1989) [3], [4], [5] examined methods for converting the speech of a male speaker to sound like that of a female speaker and vice versa. Abe et al. (1988) [6] developed a technique for voice conversion through vector quantization and spectral mapping and many more voice conversion techniques were invented during 90's and its still going on. One of the impressive work was voice conversion based on separating style and content parameter using bilinear model by Victor P et al. (2009) [7] where the author introduced text independent and dependent voice conversion.

In this work, source filter model of speech analysis and synthesis is used in which source voice (male) is changed to target voice (female) and vice versa by synthesizing excitation of source speaker with vocal tract parameters of target speaker and vice versa. Generalization is not possible but gives an idea on how voice can be transformed from one speaker to another by changing excitation and vocal tract parameters.

Organization of paper is as follows. Next section contains a brief introduction on speech synthesis using source filter model along with the algorithm for the proposed work. Section III introduces the requirement for DTW in voice

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

conversion. Section IV gives proposed voice conversion method using source filter model and section V gives conversion assessment through subjective and objective approach and finally conclusion and observations on converted speech is made.

III. SPEECH SYNTHESIS USING SOURCE FILTER MODEL

One of the most oldest and prominent method of speech analysis and synthesis is source filter model [1]. In this model speech waveform is divided into excitation parameter and vocal tract parameter in which vocal tract parameter acts as time varying all pole filter coefficients.

Speech sounds are produced by acoustical excitation of human vocal tract. The reason it is called as source filter model is because source act as excitation parameter and transfer function of vocal tract parameter acts as filter. Linear predictive coefficient as vocal tract parameter gives information on formant frequencies. Excitation of the speaker is obtained by passing the speech signal into all zero filters with LPC as filter coefficients.

Algorithm for the proposed work is as follows:

1) Get source and target speaker utterance preferably one male and one female
2) Time align both the utterances using dynamic time warping.
3) Extract LPC and excitation of source utterance and target utterance.
4) Synthesize source excitation with target LPC to get utterance in female voice.
5) Synthesize target excitation with source LPC to get utterance in male voice.

IV. DYNAMIC TIME WARPING(DTW)

In order to balance speakers speaking rate, there has to be proper time alignment between source and target speaker. Alignment can be done manually by marking the corresponding acoustical events on source and target utterance and to select the acoustical parameters which best matches between the two. But aligning manually takes much time and hence it is better to use automatic time alignment. Most preferable automatic time alignment techniques are (Hidden Markov Model) HMM based time alignment and Dynamic time warping (DTW). Figure 4a, 4b shows source and target utterance before and after time alignment respectively using DTW for the utterance “Author of the danger trail, Philip Steels, etc.” spoken by male and female speaker.

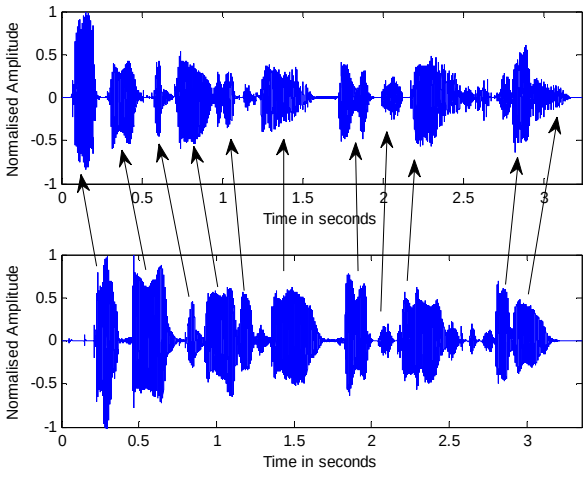

Fig 4a: source and target utterance before time alignment

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

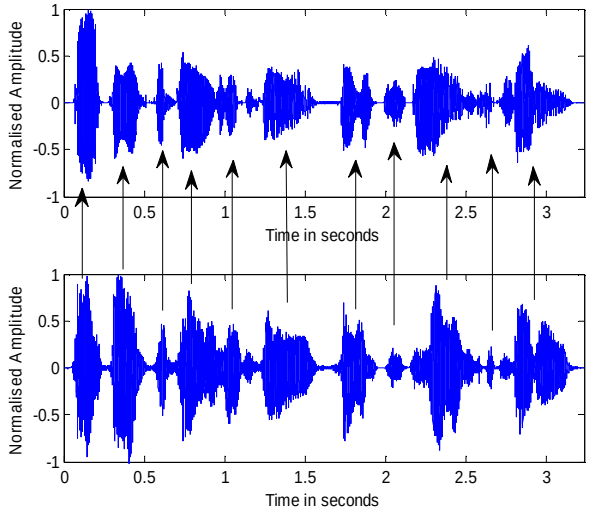

Fig 3b: source and target utterance after time alignment

V. vOICE INTERCHANE USING SOURCE FILTER MODEL

To get excitation of source and target, pass the time aligned source and target speech through FIR filter with source and target LPCs respectively as filter coefficients as shown in figure.5a below.

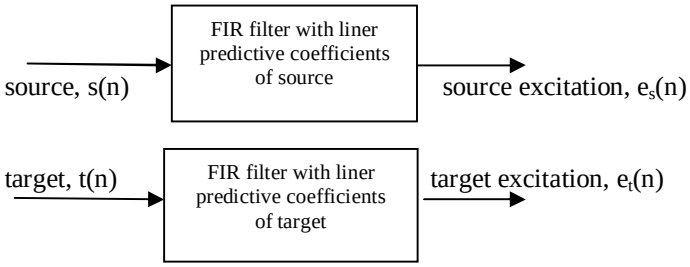

Figure.5a: extracting excitation parameters from source and target speaker

Vocal tract parameters of the source and target are extracted by windowing the time aligned signals with 25% overlap and 16 dimensional LPC was chosen. Once excitation parameters are extracted, next step is to synthesize speech by exchanging the parameters. Synthesis is done by passing excitation to the all pole filter with inverted poles of FIR filter. Figure 5b shows synthesis of excitation and vocal tract parameters and figure 4c shows utterance of original female voice and converted female voice.

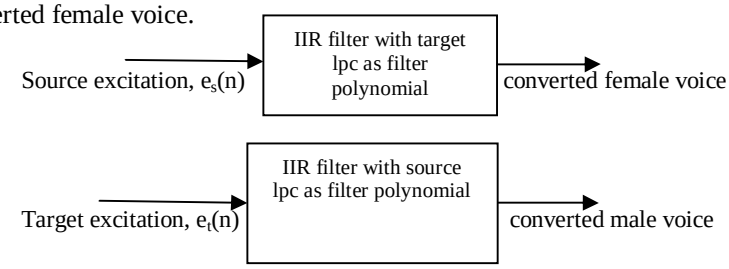

Figure.5b: synthesis of converted female and male voice.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

VI. ASSESSMENT

In order to assess the interchanged voice, mean square error (MSE) was measured between the interchanged LPC vectors and corresponding target LPC vectors. The frame wise Mean Square Error is given by[7],

$$MSE(lpc(s), lpc(t)) = \frac{\sum_{i=1}^{i=16}(lpc(s) - lpc(t))^2}{16}$$

this is averaged over entire data. In the above equation, lpc(s) and lpc(t) are the source and target lpc vectors respectively. Table. 6a below shows the MSE between interchanged and original speech which gave us a better result compared with subjective test result.

Table. 6a: MSE for the interchanged speech

Parameter	Value
MSE	0.2098

To check the correctness of interchanged voice in terms of spoken word, individual words were labeled and played separately by comparing it with the source speaker as shown in figure 6.a and 6.b

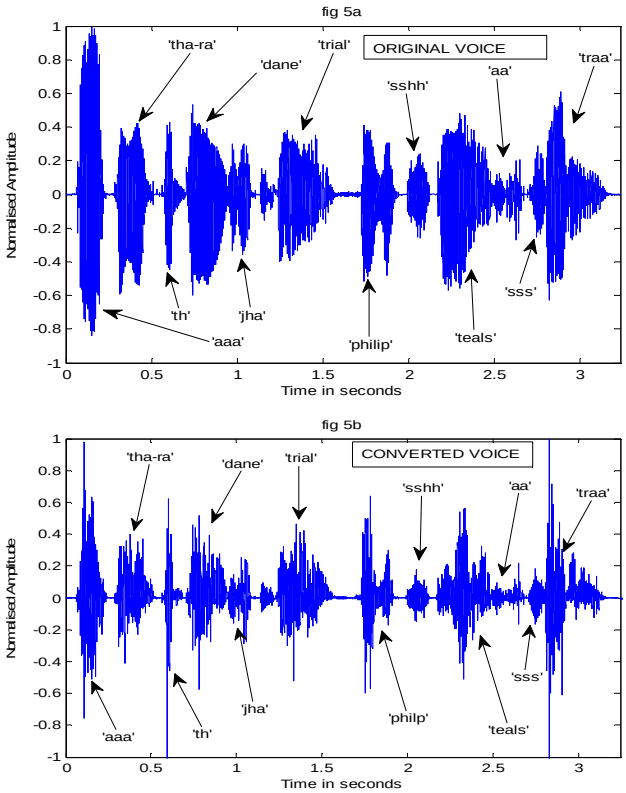

Fig 6a shows original female voice and figure 6b shows converted female voice

For subjective assessment Intelligibility, identity and quality are the three criteria chosen where 10 people were given their score from 1 to 5 with 1 being very poor and 5 being good. Table.6b shows summary of subjective analysis.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Table. 6b: summary of subjective analysis

Parameter	Score
Intelligibility	3.8 ± 0.54
Identity	3.4 ± 0.66
Quality	2 ± 0.62

VII. CONCLUSION

As a conclusion we would like to point out that even though it is possible to interchange voice characteristics by simply interchanging excitation and vocal tract parameter of speaker, we have to compromise in voice quality. One of the main reasons in quality degradation is because of synthesizing of warped speech. Since warped speech in figure 3b is degraded and we are using warped speech to synthesize the speech there is comparable amount of degradation in quality which can be observed easily in figure 4c. From subjective result it can be easily seen that there is a good score for intelligibility and identity compared to quality of voice.

Figure 7a , 7b shows the formant frequencies of first three voiced speech /aaa/, /tha-raa/ of original and converted speech and figure 7d, 7e and 7f shows plot of first formant frequency f1, second formant frequency f2 and third formant frequency f3 respectively with respect to all voiced speech in the utterance. It is observed that in terms of formant frequencies it has given close similarities which intern proves the closeness of interchanged voice.

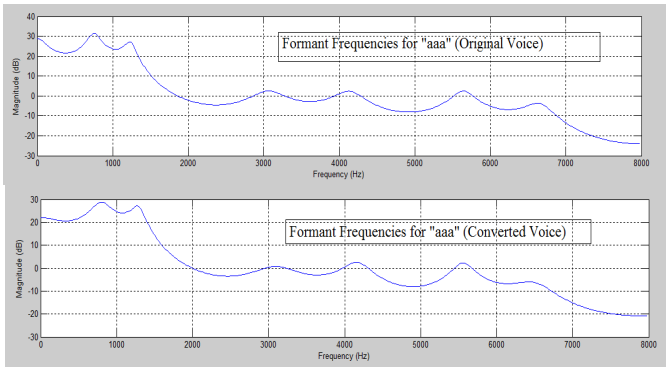

Fig 7a: formant frequencies for the voiced sound 'aaa'

Figure 6a shows the comparison of formant frequencies of original and converted speech utterance 'aaa' and it can be observed from the figure that formant frequencies are almost matching.

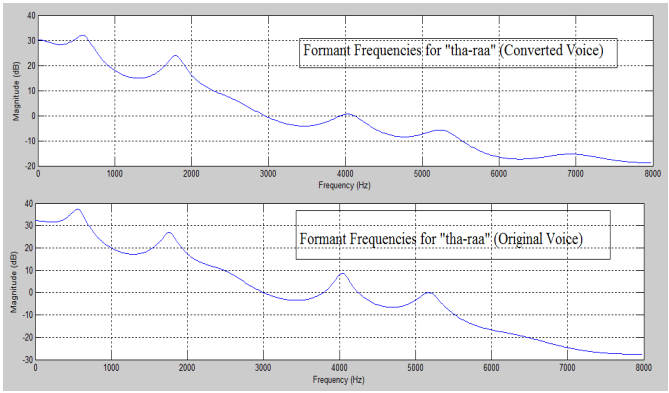

Fig 7b: formant frequencies for the voiced sound 'tha-raa'

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Figure 6a and 6b above shows the comparison of formant frequencies of original and converted speech utterance 'aaa' and 'tha-raa'. It's a plot of formant frequency versus magnitude and it can be observed from the figure that formant frequencies are almost matching to one another.

All the voiced segments are taken in to consideration with respect to first, second and third formant frequencies f1, fand f3 respectively and ploted for original and converted speech as shown in the figure 6c to 6e.

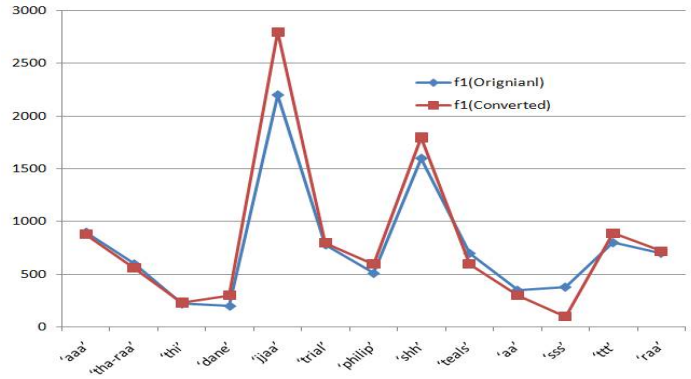

Fig 7c: Plot of voiced segments verses first formant frequency f1.

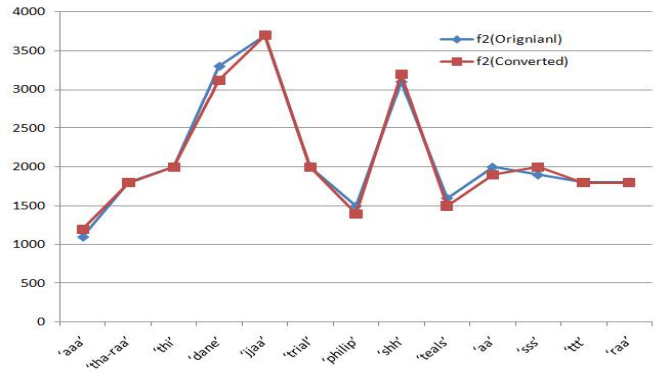

Fig 7d: Plot of voiced segments verses second formant frequency f2.

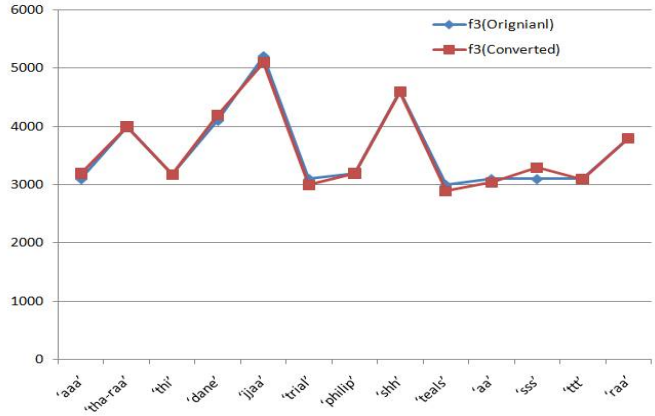

Fig 7e: Plot of voiced segments verses second formant frequency f3.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

REFERENCES

[1] B.S. Atal and S.L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Sot. Amer., Vol. 50, No. 2, pp. 637-655.
[2] S. Seneff (1982), "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-30, No. 4, pp. 566-578.
[3] D.G. Childers, B. Yegnanarayana and K. Wu (1985), "Voice conversion: Factors responsible for quality", Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 19.10.1- 19.10.4
[4] D.G. Childers, K. Wu and D.M. Hicks (1987), "Factors in voice quality: Acoustic features related to gender", Proc. IEEE Internat. ConjY Acoust. Speech Signal Process., pp. 293-296.
[5] D.G. Childers, K. Wu, D.M. Hicks and B. Yegnanarayana (1989), "Voice conversion", Speech Communication, Vol. 8, No. 2, pp. 147-158.
[6] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), "Voice conversion through vector quantization", Proc. IEEE Internat. Conf: Acoust. Speech Signal Process., pp. 655-658.
[7] V. Popa, J. Nurmien and M. Gabbouj, (2009) "A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models," Interspeech, Brighton, 6-10 September.